# 互联网可扩展路由研究

唐明董 张国清 杨景 张国强

**摘 要**: 全球路由表的高速膨胀使得当前的互联网域间路由系统的可扩展性面临着严峻的挑战。为了缩减路由表, 很多研究提出了新的路由解决方案。本文在介绍了互联网路由系统现状之后, 从较高层次上将存在的路由解决方案分为短期方案、路由架构和可扩展路由算法三部分，着重讨论了路由算法和路由架构这两类工作, 对经典的可扩展路由算法和路由架构进行了分析和比较，最后对尚未解决的问题和未来的研究方向进行了总结和展望。

**关键词**: 域间路由；可扩展性；路由算法；路由架构

如今，互联网域间路由系统的扩展性正面临着十分严峻的挑战[1]。据统计，基于 IPv4 的全球路由表(global routing table)表项数目前已在 30 万以上，且还在呈现指数级增长[2]。庞大的路由表显著增加了路由器的内存和处理器开销，导致通信延时的增长和路由收敛属性的恶化。

为了应对路由表的膨胀，网络服务提供商(ISP)采取了以下措施：一、升级路由器硬件；二、压缩路由表数据结构；三、过滤 IP 前缀。但是升级路由器硬件提高了网络服务提供商的经营成本，降低了网络的性价比，而高端路由器的性能发展能否跟上路由表的膨胀速度还是一个问题。压缩路由表的数据结构会引起更多的计算代价，不利于路由器的快速反应；过滤 IP 前缀将导致一些站点不可达。这些措施都没有触及根本问题。许多专家认为，为了从根本上解决路由扩展问题，修改边界网关协议（Border Gateway Protocol, BGP）甚至建立全新的路由架构十分必要[1]。为此，近年来针对互联网路由扩展问题的研究掀起了一个热潮，从不同技术角度对可扩展的互联网路由进行了探讨，提出了很多路由解决方案。本文对这些工作进行了综述，分析和比较了它们的基本思想和特点，并指出了有待解决的问题和未来的研究方向。

## 1 背景:互联网路由系统现状

互联网是由许多自治系统(Autonomous System，AS)连接而成的。一个自治系统可以自主决定在内部如何选择路由。网络运营商通常对自治系统内部的链路分配代价，然后沿链路代价之和最小的路径转发流量。这类路由选择协议有 OSPF[1]、IS-IS[2]等。对于较大的自治系统，它的网络通常被分为若干个路由区域以便降低路由复杂性和提高路由扩展性。

在自治系统之间唯一使用的路由协议是 BGP 协议。每个 BGP 路由器会告诉它的邻居哪些目的地址前缀标识的站点通过它的网络可达以及需要穿越的自治系统路径。因此，BGP 协议是基于路径向量的。在互联网的边缘网络中，BGP 路由器维护的路由表项数相对较少，对目的地未知的包使用缺省路由发送。然而在互联网的核心区域，BGP 路由器并不存在缺省路由，因此该区域又被称为互联网的无缺省区(default-free zone，DFZ)。无缺省区路由器常常需要为互联网的每个可达的 IP 前缀安装一条路由，结果导致路由表随着全球 IP 前缀数量的增加而膨胀。

---

1 即“Open Shortest Path First”，是一个内部网关协议
2 Intermediate system to intermediate system，中间系统到中间系统。一种内部网关协议

互联网域间路由的扩展问题早就存在。IETF[3]在上世纪 90 年代采用无类别域间路由[4]一度有效地降低了全球路由表的膨胀速度。然而,近年来各种反聚合因素的增长使得无类别域间路由的路由聚合作用逐渐失效，IPv4 前缀数量迅速增加，无缺省区的路由表再度呈现爆炸式增长。根据互联网架构委员会(IAB)在 2007 年的报告[1] ，这些因素主要包括：

1. 与提供商无关的地址

客户网络倾向于使用与提供商无关(provider-independent，PI)的地址，这样在改变提供商时可以避免重编号(renumbering)——对网络设备和主机重新分配 IP 地址。因为现实中重编号往往要花费很高的代价。PI 前缀由于不能被上级提供商聚合，必须登记到无缺省区的路由表中。增加的地址前缀不需要客户付费，然而无缺省区路由表将因此而膨胀。

2. 多宿主

多宿主(multi-homing)是指一个站点从多个提供商那里获得服务。多宿主得到广泛应用的原因在于：提供了备用路由，能够增加连接到互联网的可靠性。一个多宿主的站点可以使用 PI 地址或 PA(provider-aggregatable)[5]地址。如果使用 PI 地址，那么 PI 前缀必须出现在它的所有提供商的路由表中。如果使用 PA 地址，那么 PA 前缀仅能够被分配该地址的提供商聚合，但是不能被其它提供商聚合。实际上,由于最长前缀匹配规则的存在，为了保证 PA 前缀可以经过它的提供商可达，往往该提供商也需要单独发布该前缀。因此不管哪一种情况都将导致前缀聚合失效。

3. 流量工程

流量工程(traffic engineering, TE)的目的是让某些互联网流量避免使用特定的网络路径。使用流量工程的既包括提供商网络也包括客户网络，具体原因有负载平衡、降低费用和安全需求等。在 BGP 级，如果要对某块地址实施流量工程，那么网络运营商必须将该地址前缀从原来的较短前缀中分裂出来单独发布到全球路由表中。

上述因素使得 IPv4 前缀不断分裂，全球路由表中的前缀粒度越来越细，数量越来越多。尽管全球路由表的规模受到 IPv4 的地址空间的约束，但是这并不意味着路由表的膨胀速度会减缓。随着 IPv6 的广泛部署和应用，由于 IPv6 能够提供庞大的地址空间，可能导致全球路由表项数以更快的速度增长。基于以上因素，很多专家认为，提高互联网路由系统的扩展性已经迫在眉睫。

## 2 互联网可扩展路由研究分类

为了降低路由表规模和解决互联网路由的扩展问题，目前已经提出了许多解决方案。根据着眼点的不同这些解决方案从较高层次上可以分为三类。

1. 短期方案

短期方案大多是对 BGP 协议提出增量式的修改，并且以提高 BGP 路由的收敛属性和降低时延为主。Forgetful routing[6] [3]可以降低路由表所占用的内存空间。它的基本思想是选择

[3] Internet Engineering Task Force，互联网工程任务组

[4] Classless Inter-Domain Routing，CIDR。一个用于给用户分配 IP 地址以及在互联网上有效地路由 IP 数据包的对 IP 地址进行归类的方法

[5] 可由提供商聚合的地址

[6] 有译成“健忘路由”

性地丢弃路由表中的部分替代路由，只有在必要时才从邻接路由器那里获取。它不需要改变 BGP 协议且基本不影响路由的收敛属性。但是 Forgetful routing 并没有减少 IP 前缀数量和路由表的增长速度，因此是一个短期方案。考虑到短期方案并不能真正提高路由扩展性，本文中不作具体讨论。

2. 路由架构

从中长期来解决路由扩展性的目标出发，很多研究人员提出了新的路由架构，其中许多是 IRTF[7] [8]的提案。绝大多数的新路由架构都是考虑在现有域间路由系统中增加一个间接的中间层(indirection layer)，使域间的扁平路由架构变成分层的。中间层将 IP 地址空间分离为主机标识和路由标识两部分，后者在路由时作为位置符使用，并建立它们之间的映射；或者更激进一些，中间层可能引入一种新的名字空间作为路由标识空间，将 IP 地址空间映射到该名字空间。包在传递过程中由中间层在某个阶段将主机标识用路由标识替换来穿越互联网。路由标识要么是能够聚合的，要么具有较大的粒度，使得路由标识的数量是可控的，从而缩减路由表。

基于路由标识的类型和来源，可以将新的路由架构分成核心-边缘分离、位置符/标识符分离、基于自治系统号的路由、基于虚拟聚合对象的路由等；根据路由架构以改变主机为主还是以改变网络为主可以分为基于主机的和基于网络的；而根据包传递过程中使用路由标识替换主机标识时是改写包地址还是对包进行封装，可以将新路由架构分为地址重写(address rewriting)的和映射与封装的(map & encapsulation)。

3. 可扩展路由算法

可扩展路由算法研究旨在降低网络节点维护的状态数量和控制开销，立足于从根本上来解决路由扩展问题。针对互联网的可扩展路由算法研究一直没有停止过。最早是克兰洛克（L. Kleinrock）和卡蒙（F. Kamoun）[4]提出的基于分区的层次化路由算法。它在互联网上得到了广泛应用，如划分自治系统，以及 OSPF、IS-IS 等支持分区的域内路由协议。 而 BGP 协议采用的路由算法是基于路径向量的，本质上不具备良好的扩展性。到目前为止,已存在很多种可扩展路由算法，如不同类型的层次化路由、地理路由、紧凑路由、分布式哈希表路由等等。这些路由算法有的被互联网路由系统所接受，如基于提供商的层次化路由，更多的还停留在理论层面上，作为互联网长期可扩展的新型路由方案的探索。

瑞格特（Yakov Rekhter）有一句名言：“要么编址服从拓扑，要么拓扑服从编址，二者必居其一（Addressing can follow topology or topology can follow addressing, choose one）。”这一针对路由扩展性的基本假设被称为 Rekhter 定律。针对互联网的可扩展路由研究大都是试图重新将互联网路由扳回到遵循该定律的轨道上来。

# 3 可扩展路由研究成果分析

## 3.1 可扩展路由算法

传统的最短路径路由算法,如距离向量、链路状态和路径向量等算法，在每个节点上要维护到所有节点的路由信息，因此路由表项数为 $O(n)$，其中 $n$ 是节点总数。此时路由表占用的内存和维护路由表的控制开销都随网络规模快速增长，因此扩展性不好。可扩展路由算法旨在降低路由表的内存开销和控制消息开销。下面分类介绍已知的一些经典的可扩展路由

[7] Internet Research Task Force，互联网研究任务组

算法。

### 3.1.1 基于分区的层次化路由

基于分区的层次化路由(area hierarchical routing)的基本思想是：对网络嵌套地划分区域并隔离不同区域的拓扑更新，每个节点对自己维护的路由信息采用“距离越近越详细”的策略，较远的区域信息尽量简略。基于分区的层次化路由是由克兰洛克和卡蒙[4]最早提出的，也称基于分簇的层次化路由(cluster hierarchical routing)。这一路由方式可以使节点的路由表项数缩减至 $kn^{1/k}$ ，其中 $n$ 是节点总数， $k$ 是分区的级数。

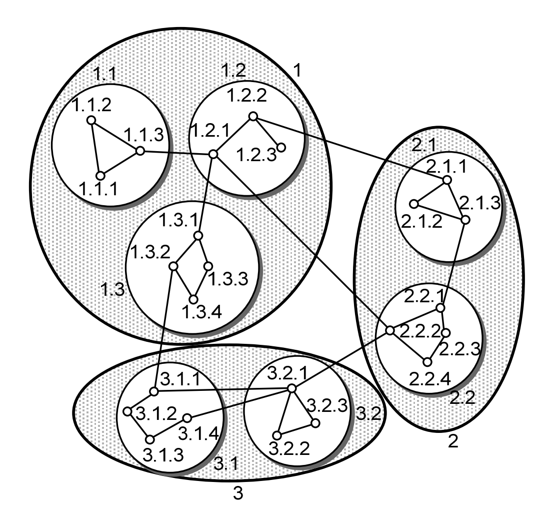
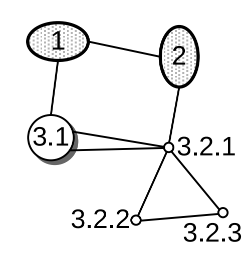



**图1.** 基于分区的 3 级层次化网络

下面简单介绍这种路由的基本设计。 图 1 给出了一个具有 3 级的层次化网络的例子。 图 2 则示出了图 1 中节点 3.2.1 所能见到的网络视图和路由表设置。节点 3.2.1 的路由表项数总共为 6(而如果不划分区域,需要 24 项，等于网络节点总数)。




| 目的地 | 下一跳 | 跳数 |
|---|---|---|
| 1 | 3.1.1 | 2 |
| 2 | 3.2.2 | 1 |
| 3.1 | 3.1.1 | 1 |
| 3.2.1 | —— | 0 |
| 3.2.2 | 3.2.2 | 1 |
| 3.2.3 | 3.2.3 | 1 |

**图2.** 节点 3.2.1 的网络视图和路由表项

基于分区的层次化路由思想在互联网上得到了广泛研究和应用。例如，互联网划分为不同的自治系统，使用域间——域内两级的路由结构；并且在域内，OSPF 和 IS-IS 协议也是分区的。针对域间路由，最近的一些研究也提出了分区的层次化路由架构[5]。

### 3.1.2 基于地标的层次化路由

基于地标的层次化路由(landmark hierarchical routing)是对基于分区的层次化路由的改编，以使层次结构更易于动态管理。它最早是由土屋（Paul Tsuchiya）[6]提出的，基本思想是：迭代地选择网络中的节点作为地标并构建层次结构，每个节点只存放到本地节点和若干地标的路由信息；路由时对目的地不在路由表中的消息朝离目的节点最近的地标发送，然后由地标转发给目的节点。

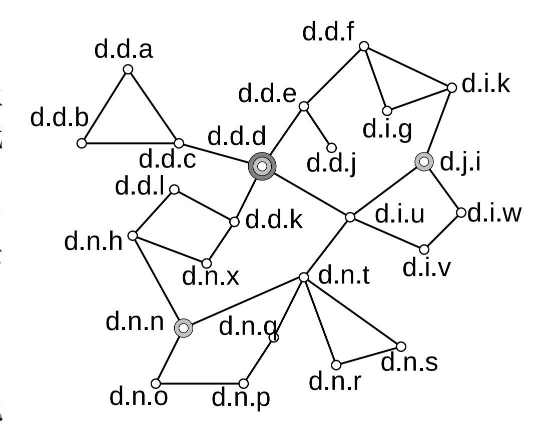



**图3.** 基于地标的 3 级层次化网络

这类路由的一个重要概念就是地标。一个地标是从网络中选择的一个节点。设该地标的管辖半径为 r，那么可以将与它的距离不超过 r 的那些节点的集合看成是它的邻域，邻域中的每个节点都包含到该地标的路由信息。值得注意的是，地标的路由表可以不需要维护到邻域中节点的路由信息。图 3 展示了一个基于地标的 3 级层次化网络，其中所有节点都是第 0 级地标，用两个圆环表示的节点是从第 0 级地标中选出的第 1 级地标，用三个圆环表示的节点是从第 1 级地标中选出的第 2 级地标。

基于地标的层次化路由可以大幅缩减节点的路由表，典型情况可以缩减至 $O(\sqrt{n})$ [6]。基于地标的层次化路由可以克服基于分区的层次化路由存在的边界效应，邻近节点之间的路径质量得到了提高。由于具有相对较好的动态管理特性，基于地标的层次化路由比较适用于一些拓扑变化较快的网络，如无线自组织网络。

### 3.1.3 基于提供商的层次化路由

基于提供商的层次化路由(provider hierarchical routing)是目前互联网唯一使用的域间可扩展路由机制，无类别域间路由的地址聚合就是以该方法为基础的。基于提供商的层次化路由的基本思想是利用域间存在的提供商-客户(provider-customer)层次结构对客户网络分配可由提供商聚合的 IP 地址（PA 地址）。

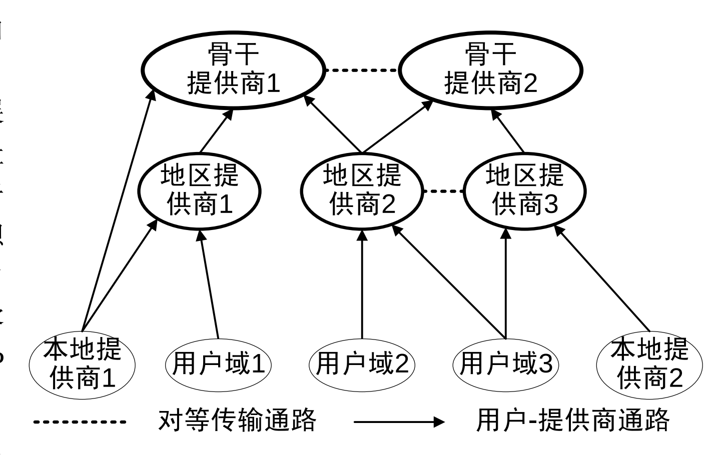



图4. 基于提供商的互联网层次结构

图 4 展示了互联网的基于提供商的层次结构。用户域（Stub）必须连接到一个提供商域来获取转发(transit)服务。而本地级或地区级提供商必须连接到骨干提供商来获取转发服务。互联网的层次结构并不是树状的。一个域可以连接到多个提供商，这种情况称为多宿主。并且两个域之间还可以建立对等传输（peer to peer）关系，以互相提供穿越服务，如地区提供商 2 和地区提供商 3。一个域甚至可以直接与不同层次的提供商相连，如地区提供商 1。因此，一个域可能具有由不同的提供商分配的多个地址前缀。如果一个客户网络改变了它的提供商，那么它应当从其新提供商那里获得地址，并将原来的地址还给以前的提供商，该过程称为重编号。

网络重编号需要很高的代价。并且使用可由提供商聚合（PA）地址也将阻碍多宿主、主机移动和流量工程的应用。因此客户网络越来越倾向于使用与提供商无关（PI）地址，而不是可由提供商聚合地址。这些使得基于提供商的层次化路由在互联网上失效，从而导致路由表膨胀。

### 3.1.4 分布式哈希表路由

分布式哈希表(Distributed Hash Table，DHT)技术在网络上构建了一种键—值(key-value)查询服务。节点和键共用一个特定的标识空间。假定键对应的值存放在标识与键相等的节点上，从一个节点上根据键查找值的过程等同于从该节点到相应目的节点的路由过程。到目前为止，分布式哈希表主要用于在网络层之上构建逻辑网络，前提是存在网络层路由协议。这种逻辑网络是根据节点的标识度量空间来构建的，节点之间的逻辑距离可以由它们的标识推算出。每个节点将收到的消息转发给离目的节点“距离”最近的邻居，从而逐步逼近目的节点。分布式哈希表网络通常具有比较规则的拓扑结构，如环状[7]、树状[8]等。

凯撒（M. Caesar）等人[9]考虑了在物理网络上直接构建分布式哈希表，提出了一种虚拟环路由方法(Virtual Ring Routing，VRR)。虚拟环路由使用不包含语义的扁平标识路由，只要保证标识唯一即可。基本思想是：将所有节点根据它们的标识顺序组织成一个虚拟环，利

用类似于 Chord[8] [7]的机制来贪心地转发消息。虚拟环路由方法在多种类型的网络上的应用都得到了研究，如无线自组织网络、企业局域网和互联网等。

ROFL[9] [10]是一种基于虚拟环路由的面向互联网的可扩展路由算法。ROFL 在网络层消除了位置符，完全使用与位置无关的标识进行路由。该方法继承了位置符/标识符分离的诸多优点，如移动性、多宿主和标识稳定等，但是不需要引入映射系统，编址和路由显得更加简单。ROFL 的基本原理是在网络层将主机和路由器标识组织成层次化的分布式哈希表，同时支持基本的域间路由策略。

ROFL 能够大幅缩减路由表规模,仿真表明在 $6\times10^8$ 个 ID 的网络中，平均情况下路由表规模仅为 $10^2$ 的数量级。但是 ROFL 使用的路径与最短路径相比长度显著增加，表明在路径质量上还不太理想。

### 3.1.5 紧凑路由

紧凑路由(compact routing)是一类在理论上保证同时具有较小路由表和较低拉伸系数的路由算法。紧凑路由使每个节点的路由表规模为 $O(n)$,即随节点总数 $n$ 呈亚线性增长,因此保证了路由表的可扩展性。它的基本思想是平衡路由表大小和路径长度，允许有限的路径拉伸来大幅缩减路由表。

针对通用图的紧凑路由研究目前已取得接近最优的成果[11,12]。但是真实的网络往往呈现特定的拓扑结构特征。通用的紧凑路由算法因为没有利用拓扑特征在实际网络上可能并不是最优的。研究人员常常用特定结构的图来为真实的网络建模，如幂律图(power-law graph)、增长受限的图(growth-bounded graph)、平面图等。

将紧凑路由用于互联网域间的设想始于克留科夫（D. Krioukov）等人[13]，他们使用 TZ[10] 算法在幂律图和真实的互联网自治系统图上进行了仿真，发现平均的路由表项数很小，而平均拉伸度约为 1.1.。此后，针对类互联网拓扑结构的图上的紧凑路由开始得到更多关注。文献[14-16]等利用了无标度特征来设计紧凑路由算法，被证明比 TZ 算法在类互联网的网络上具有更高的性能。文献[17]对紧凑路由和在互联网上的应用问题进行了综述。

### 3.1.6 地理路由

地理路由(geographical routing)是指基于节点的地理位置使用贪心路由。一个节点通常只需要维护自己和邻居的地理位置，收到消息时选择在地理位置上离目的节点最近的邻居作为下一跳。地理路由有可能陷入局部最小点(local minimum)——找不到比自己离目的节点更近的邻居，因此贪心路由失效。解决该问题的常用办法是使用基于拓扑的替代路由方案，如面路由[18]等。一旦贪心路由遇到局部最小点时，启用替代路由方案来跳出局部最小点，然后再恢复贪心路由。

地理路由可以在每个节点上使用很小的路由表，几乎不需要维护网络的拓扑信息，因此具有十分理想的可扩展性。地理路由在互联网和无线网络领域都得到了广泛研究。文献[19]较早提出了在互联网上如何使用地理信息编址和路由，并与基于提供商的路由方案进行了比较；文献[20]针对 IPv6 提出了使用基于地理位置的编址方案；文献[21]则考虑了在包首部中携带地理信息来辅助互联网路由。

---

8 一个分布式查找协议

9 Routing on flat labels，一种完全基于平面标签的新颖路由方式

10 Thorup-Zwick

为了保证较高的路由成功率和较低的拉伸度，地理路由要求网络连接密度尽可能高，即要求地理位置靠近的节点在拓扑上也是邻近的，最好是直接互连的。然而，真实网络可能难以满足上述条件。特别是在互联网环境中，自治系统之间的连接关系是由它们之间的利益决定的，因此相邻地区的自治系统并不一定是互连的；况且地理路由在支持流量工程方面还存在问题。

### 3.1.7 图嵌入路由

图嵌入(graph embedding)的基本思想是：对网络中的每个节点分配虚拟坐标，将节点映射到虚拟几何空间或隐藏度量空间中的点，简化节点之间的“距离”计算。通过图嵌入，路由算法也可以使用贪心路由，每个节点只需要知道邻居的坐标，在转发消息时总是选择与目的节点“距离”最近的邻居转发，这一点与地理路由相似。但是与地理路由方法不同，虚拟坐标的构造不需要感知节点的物理位置，而是通常基于网络的拓扑信息构造。这样做的优点是虚拟坐标能够反映网络的连通信息，在很多情况下能够提高路由的性能。缺点是当网络的拓扑改变时，至少一部分节点的虚拟坐标也要随之更新，因此动态属性不如地理路由。

基于图嵌入的路由研究目前大多针对无线网络[22-24]，但是可以为未来的互联网路由设计提供启发。克留科夫等人[25]认为基于复杂网络的隐藏度量空间的贪心路由有可能为互联网提供一种理想的路由方法。 [25]提出了利用一种双曲空间来构造类互联网的无标度网络的方法，发现基于该双曲空间的贪心路由具有较为理想的性能。[26]提出了将无标度网络嵌入到由它的骨架导出的度量空间的方法，分析表明基于该度量空间的贪心路由具有高扩展性。

## 3.2 路由架构

为了克服现有域间路由架构的不足,很多研究提出了新的域间路由架构，其中许多是 IRTF 的提案。这些路由架构依据基本特点可以分为成下面几类。值得注意的是，该分类并不是严格的划分，一种路由架构可能兼有多种特点，因此亦可属于不同类别。

### 3.2.1 基于自治系统号的路由

由于自治系统具有比 IP 前缀更大的粒度，因此一些研究提出在域间基于自治系统号而不是 IP 前缀来进行路由。这样，核心路由表的表项数最多等于自治系统的总数。由于自治系统总数目前比 IP 前缀总数少一个数量级，因此可以大幅缩减路由表规模。这类工作的一个代表是 HLP[27]。HLP 还利用了划分区域的分级路由思想。它根据服务提供商-客户（provider-customer）的层次关系将域间网络划分成树状区域，把不同的区域隔离，将路由更新和故障限制在本区域内。HLP 使用了链路状态和路径向量两种路由协议来提高域间路由的收敛属性。但是域间网络拓扑由于多宿主和对等操作（peering）[11]连接的广泛使用，已经远离树状结构，划分树状区域在效果上并不理想。而且 HLP 要求网络服务提供商公开它们之间的连接关系，这在目前也是不太可行的。

原子路由（Atomized routing） [28] 引入了一种称为“原子”的对象，聚合那些拥有相同自治系统路径的长 IP 前缀，在一定程度上也可以看成是基于自治系统号的路由。原子路由被设计用于边缘的客户网络，即被通告的原子来自于客户网络，在核心网络中根据原子标识或 IP 前缀进行路由。通过用原子标识聚集长 IP 前缀，使核心路由表规模得到了缩减。

基于自治系统号的路由必须构建和维护 IP 前缀到自治系统号的映射表，并提供查询。从长期来看，这类路由架构的路由扩展性仍然受到较强的限制。这是因为近年来分配的自治

[11] 网络服务商之间的一种数据流流通安排：在两个服务商之间交换路由通告，以确保来自第一个服务商的业务能够到达第二个服务商的客户，反之亦然。

系统号的增长速度比 IP 前缀数量增长更快，从 2000 年的 10000 个左右迅速增加到目前的近 50000 个[2]，自治系统号目前已从 16 位升级到 32 位。照此趋势，全球路由表的增长速度仍得不到有效控制。

### 3.2.2 基于虚拟聚集对象的路由

这类路由架构也是基于比 IP 前缀更大的拓扑粒度来路由。基本思想是使用虚拟聚集对象来聚集较长的 IP 前缀，在核心网络中使用虚拟聚集对象的标识来路由。这类工作主要有 CRIO[29]、ISLAY[30]等。

CRIO 使用一种虚拟 IP 前缀对实际的 IP 前缀进行聚合。虚拟前缀与核心区域的路由器关联。这些路由器称为聚合代理(aggregation proxy)。聚合代理用较短的前缀能够聚合很多较长的 IP 前缀。聚合代理向其它核心路由器发布自己的虚拟前缀，并负责将接收到的包用隧道方式路由到与实际的 IP 前缀对应的路由器。由于核心区域使用了虚拟前缀，CRIO 方法可以使核心路由表缩小两个数量级，但是路径长度比原来的路径有所拉伸。实际上,虚拟前缀的粒度越大，路径的拉伸度越高，因此存在一个利弊平衡问题。CRIO 与原子路由有很多相似之处。但是 CRIO 的虚拟前缀分配可以与网络拓扑无关，并且 CRIO 将所有的改变限制在运营商网络中。

ISLAY 使用的虚拟聚集对象称为“aggregate”,相应的标识称为 aggregate ID。一个 aggregate 可以聚合多个其它 aggregate 和 IP 前缀。核心路由表中仅使用 aggregate ID，同样大幅缩小了路由表规模。

这类路由架构也需要建立和维护 IP 前缀到虚拟聚集对象标识的映射表，并提供查询。

### 3.2.3 核心与边缘分离

核心—边缘分离(core-edge separation)架构的基本思想是将提供商网络(即转发网络)与边缘的客户网络(即 stub 网络)的地址和路由空间隔离。这种架构的提出者认为，互联网路由扩展问题的根源在于核心网络和边缘网络使用同一个地址和路由空间，而核心网络的地址聚合要求与客户网络的流量工程、多宿主等反聚合因素存在矛盾。核心网络无论是在 IP 前缀数量上还是拓扑上都相对比较稳定，而边缘网络的 IP 前缀数量和引发的路由更新都要多得多，并且增长更快。通过隔离核心网络和边缘网络，在核心路由器中只维护核心网络通告的 IP 前缀，可以大幅缩小核心路由表，并且能够隔离客户网络引起的路由更新，减少路由更新数量和提高路由收敛属性。这方面的代表工作如 eFit[5]。

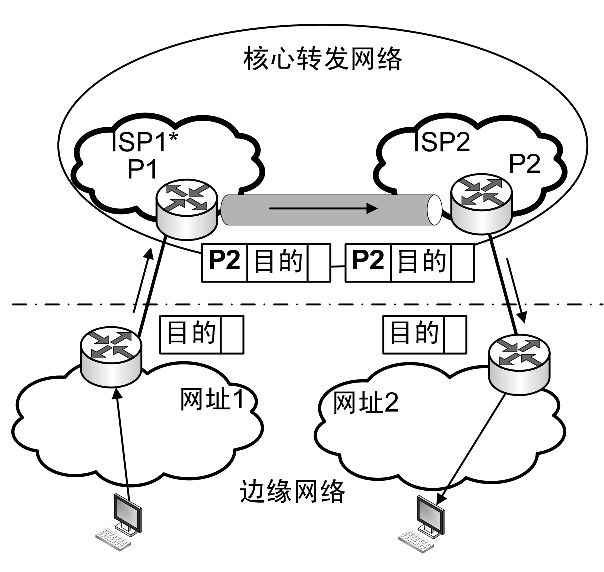



图5. 核心网络与边缘网络隔离的路由架构

eFit 提出将 IP 地址空间划分为两种，即提供商地址和用户地址空间。前者在核心网络中使用，而后者针对边缘网络，不能出现在核心路由表中，但必须是全局可达的。图 5 给出

了 eFit 架构中端到端的路由过程：源自客户站点的 IP 包先使用目的用户地址路由到提供商网络的入口边界路由器 P1；然后 P1 根据用户地址通过映射系统查询对应的目的提供商地址，即与目的站点相连的提供商网络的边界路由器的地址，设为 P2；在 P1 处将 P2 作为首部对 IP 包进行封装，将包隧道至 P2，P2 对收到的包进行拆封，再使用目的用户地址路由至目的主机。eFit 对核心路由表的缩减程度依赖于提供商地址的聚合程度。因此核心路由表规模大体上与提供商网络数量呈线性比例。

核心—边缘分离的思想实际上在多数路由架构和提案中都得到了不同程度的体现。不同的是，其它的路由架构，如位置符与标识符分离等，并不严格要求隔离边缘和核心的路由空间，而是允许路由协议穿越边界。核心—边缘分离方案也需要一个映射系统来建立和维护用户地址空间到提供商地址空间的映射，并提供查询功能。

### 3.2.4 位置符与标识符分离

位置符/标识符分离 (Locator/ID separation)架构的提出者认为现有互联网路由系统扩展性差的一个因素是 IP 地址语义过载，即既作为位置符又作为标识符。当 IP 地址标识终端身份时，它是根据网络服务提供商组织结构而非拓扑结构分配，导致目前互联网唯一有效的路由缩减技术——拓扑聚合的失效。解决方法是将单一的 IP 地址空间分离为位置符与标识符两种地址空间。位置符是基于拓扑分配的，可由提供商聚合，而标识符不需要依赖于拓扑，可以在应用层和传输层使用。因此，拓扑变化时标识符不需要改变，只需要改变相关的位置符。值得说明的是，位置符与标识符分离方案强调尽可能保持互联网原有的路由协议不变，因此可部署性相对较好，是受到最多关注的一类路由架构。最近几年提出的这类路由架构和 IRTF 提案有 LSIP[31]、Ivip[32]、Six/One Router[33]等。思科（Cisco）公司已经实现了 LISP 的一个原型系统。

位置符和标识符分离本身并不能提高路由的扩展性，关键是位置符/标识符分离的位置。多数方案都是在互联网核心—边缘处分离位置符与标识符。位置符分配给核心网络和边缘网络的网关(路由器)，而标识符分配给客户网络和终端设备。一般在客户网络本地可以使用标识符交付包，但是穿越核心网络通常使用隧道方式，即在隧道入口路由器(ITR)通过映射系统查询目的站点的位置符，然后使用目的位置符在核心网络中路由，到达隧道出口路由器(ETR)，然后再使用目的主机的标识符路由。ITR 和 ETR 可以放置在客户网络中，也可以放置在提供商网络中。

利用位置符与标识符分离，客户网络完全可以使用与提供商无关的 IP 地址，在改变提供商时不需要重新编号。通过使用与拓扑无关的标识符，这类方案对客户网络在流量工程、多宿主和移动性等方面的扩展性也提供了良好的支持。而对于提供商网络，由于位置符可以按基于提供商聚合的方式分配，因此核心路由表规模能够得到大幅缩小。但是提供商网络的流量工程将受到位置符聚合的约束。

### 3.2.5 支持源路由的架构

这类路由架构的特点主要是在解决路由扩展性问题的同时支持用户源路由选择和多路径路由。代表工作有 NIRA[34]和 Pathlet Routing[35]。

NIRA 根据用户站点到互联网核心(通常由 tier-1 自治系统构成)或对等连接点的上行路径来对用户站点编址，一条路径用一个地址来编码，这种编址方式是基于提供商聚合的。NIRA 主要是针对 IPv6 的，但是也可以改编用于 IPv4，此时一条上行路径可以用一串 IPv4 地址来编码。假设使用 IPv6 地址，那么从源主机到目的主机的一条路径可以通过它们的地

址来联合推出。如图 6 所示，网络 N1 的用户 A 具有从核心网络分配来的地址 1 :1 :1 ::1000 和 1 :2 :1 ::1000，以及从对等连接点分配来的地址 FFFF :1 :1::1000。 网络 N3 的用户 B 具有从核心网络分配来的地址 2 :1 :1 ::2000 和 1 :3 :1 ::2000，以及从对等连接点分配来的地址 FFFF :2 :1::2000。 用户可以使用源地址和目的地址来计算一条合法的路径。如用户 A 可以使用源地址 1 :1 :1 ::1000 和 B 的地址 1 :3 :1 ::2000 来得到一条路径 N1→R1→B1→R3→N3，或者用源地址 FFFF :1 :1::1000 和 B 的地址 FFFF :2 :1::2000 得到路径 N1→R2→R3→N3。

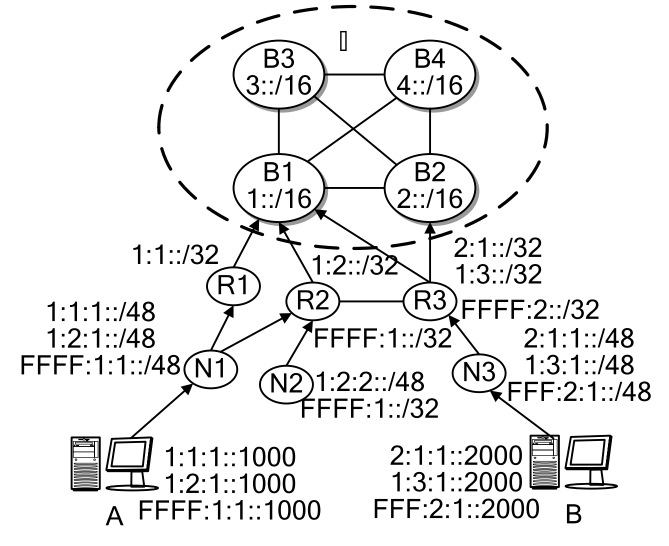



**图6.** 一种支持用户选择路由的架构——NIRA

Pathlet 路由的基本想法是每个自治系统中的路由器聚合成一些抽象的虚拟节点 vnode，自治系统愿意提供的路由线路就是由一串 vnode 构成。这样的 vnode 序列称为 pathlet。每个自治系统向互联网发布自己提供的 pathlets。源站点可以将连接源到目的地的一串首尾相连的 pathlets 作为一条端到端的源路由，从而实现多路径路由。这个框架提供了比较灵活的路由选择，因为 vnodes 能够以任意的粒度来抽象网络，pathlets 能够表达路由的约束策略，同时提供了大量可选的路径，从而使得源站点能够按指数级的组合方式缝合 pathlets，形成端到端路由。Pathlet 路由结合了 BGP 和源路由这两种路由策略的优点，其代价是更复杂的控制平面和更多的控制消息开销。

## 4　总结与展望

虽然针对互联网的可扩展路由研究取得了很多成果，但是还没有一种方案得到应用和部署。IRTF 的路由工作组还在不断讨论收到的路由架构提案。新的路由解决方案还在不断形成过程中。鉴于互联网的重要性和复杂性，对现有互联网路由系统做出改变并不是一件容易的事情，有很多问题需要考虑和解决。在已有的互联网可扩展路由研究的基础上，我们对未来需要解决的问题和研究方向进行了归纳，分为路由架构和路由算法两个方面。

从路由架构来看，位置符/标识符分离得到了较多的关注和认同。除了可以解决路由扩展问题外，这种架构还带来了许多额外的好处，能够更好地满足当前路由架构无法支持的应用需求，例如流量工程、终端移动等。但是还存在以下的问题：

(1). 位置符/标识符分离是否真正最好的路由架构？位置符/标识符分离方案依赖于隧道技术，创建了一个全然不同的网络。然而隧道技术增加了开销并提高了网络延迟，有时甚至会因为传送数据过大而丢失数据。因此仍有必要继续探索新的更好的路由解决方案；

(2). 位置符/标识符分离架构中，如何构建最好的映射系统？映射系统必须具有高扩展、安全性和低延迟等性质。映射系统设计涉及：映射表建立机制，映射表项更新、撤销机制，映射表查询方法，映射表服务器的部署、管理等。如何设计和评估映射系统仍将是未来的一个研究热点；

(3). 位置符/标识符分离架构中，包封装方法是否真正优于地址重写方法，还需要研究和评

估；

(4). 如何平衡路由表扩展性和日益增长的流量工程需求？位置符/标识符分离的架构依赖于位置符的聚合来缩减路由表，然而流量工程要求使用更细的地址前缀，如何达到最好的平衡还需要研究;

(5). 对映射系统引入的利益关系，即谁付费、谁受益还需要有更深入的理解。比如映射系统谁来部署，是否需要向用户收费等。

另一方面，面向未来互联网的可扩展路由理论仍然值得进一步研究。这方面的工作目前聚焦在寻求比地址聚合更好的可扩展路由算法，如紧凑路由、贪心路由等。地址聚合是当前互联网唯一使用的可扩展路由机制。然而互联网自治系统级拓扑结构逐渐从树状演化成网状（mesh），呈现无标度特征。一些专家认为，地址聚合在本质上越来越与域间拓扑结构背离，故从长期来看，扩展性将受到限制。因此有必要充分考虑互联网的结构和演化特性来设计高效的路由算法，为新的可扩展路由方案提供理论与实践指导。

**参考文献：**

[1] Meyer D, Zhang L, Fall K, Report from the IAB workshop on routing and addressing, *The Internet Architecture Board*, 2007.

[2] BGP Routing Table Analysis Reports, *http://bgp.potaroo.net/*.

[3] Karpilovsky E and Rexford J, Using forgetful routing to control BGP table size, In *CoNEXT, 2006*.

[4] Kleinrock L, Kamoun F, Hierarchical routing for large networks: Performance evaluation and optimization, *Computer Networks*, 1:155–174, 1977.

[5] Massey D, Wang L, Zhang B, and Zhang L, A proposal for scalable Internet routing and addressing, Internet Draft, http://www.ietf.org/internet-drafts/draft-wang- ietf-efit-00.txt, 2007, 2.

[6] Tsuchiya P, The landmark hierarchy: A new hierarchy for routing in very large networks, In *Proceedings of ACM SIGCOMM*, 1988.

[7] Stoica I, Morris R, Karger D, Kaashoek MF, and Balakrishnan H, Chord: A scalable peer-to-peer lookup service for internet applications, In *Proceedings of ACM SIGCOMM*, San Diego, California, August 2001.

[8] Plaxton CG, Rajaraman R, Richa AW, Accessing nearby copies of replicated objects in a distributed environment, In *Proceedings of the ninth annual ACM symposium on Parallel algorithms and architectures*, pages 311–320, Newport, Rhode Island, United States, 1997.

[9] Caesar M, Castro M, Nightingale E, O'Shea G, Rowstron A, Virtual ring routing: network routing inspired by DHTs, In *Proceedings of ACM SIGCOMM*, 2006.

[10] Caesar M, Condie T, Kannan J, Lakshminarayanan K, Stoica I, and Shenker S, ROFL: Routing on flat labels, In *Proceedings of ACM SIGCOMM*, 2006.

[11] Thorup M, Zwick U, Compact routing schemes, In *Proceedings of the 13th ACM Symposium on Parallel Algorithms and Architecture (SPAA 2001)*, Heraklion :ACM Press, 2001,1-10.

[12] Abraham I, Gavoille C, Malkhi D, Nisan N, and Thorup M, Compact name-independent routing with minimum stretch, In *SPAA*, 2004.

[13] Krioukov D, Fall K, Yang X, Compact routing on Internet-like graphs, *In Proceedings of INFOCOM*, Hong Kong IEEE, 2004, 209-219.

[14] Enachescu M,Wang M,Goel A, Reducing maximum stretch in compact routing, *In Proceedings* of INFOCOM, Phoenix:IEEE, 2008, 977-985.

[15] Brady A, Cowen L, Compact routing on power-law graphs with additive stretch, In *Proc. of the 8th Workshop on Algorithm Engineering and Experiments*, 2006, 119–128.

[16] 唐明董,张国清,杨景,张国强, 针对无标度网络的紧凑路由方法, 《软件学报》, 已录用待发表.

[17] Krioukov D, Fall K, Brady A, On compact routing for the Internet, *ACM SIGCOMM Computer*

*Communication Review*, 2007(7), 37:43-52.

[18] Karp B and Kung H.T., GPSR: greedy perimeter stateless routing for wireless networks, In *Proceedings of the 6th Annual MOBICOM*, *ACM Press*, 2000, pp. 243-254.

[19] Francis P, Comparison of geographical and provider-rooted Internet addressing, *Computer Networks and ISDN Systems*, 27(3):437–448, 1994.

[20] Hain T, An IPv6 provider-independent global unicast address format, *Internet Draft*, *http://tools.ietf.org/html/draft- hain-ipv6-pi-addr-10*, 2006.

[21] Oliveira R, Lad M, Zhang B, Zhang L, Geographically informed inter-Domain routing, *ICNP*, 2007.

[22] Fonseca R, Ratnasamy S, Zhao J, Ee CT, Culler D, Shenker S and Stoica I, Beacon vector routing: Scalable point to point routing in wireless sensornets, In *Proceedings of the Second USENIX/ACM Syposium on Networked Systems Design and Implementation (NSDI 2005)*, 2005.

[23] Kleinberg R, Geographic routing using hyperbolic space, In *Proc. IEEE INFOCOM*, 2007.

[24] Flury R, Pemmaraju S, and Wattenhofer R, Greedy routing with bounded stretch, In *Proc. IEEE INFOCOM*, 2009.

[25] Krioukov D, Papadopoulos F, Boguna M, and Vahdat A, Greedy forwarding in scale-free networks embedded in hyperbolic metric spaces, *ACM SIGMETRICS Performance Evaluation Review*, Vol. 37, no.2, p15-17, 2009.

[26] 唐明董,张国清,杨景，大规模网络上基于图嵌入的可扩展路由方法研究，《计算机研究与发展》，已录用待发表.

[27] Subramanian L, Caesar M, Ee C.T., Handley M, Mao M, Shenker S, Stoica I, HLP: A next-generation Interdomain routing protocol, In *Proceedings of ACM SIGCOMM*, August 2005.

[28] Verkaik P, Broido A, claffy kc, Gao R, Hyun Y, Ronald van der Pol, Beyond CIDR aggregation. *Technical Report TR-2004-1,CAIDA*, 2004.

[29] Zhang X, Francis P, Wang J and Yoshida K, Scaling IP routing with the core router-integrated overlay, *IEEE International Conference on Network Protocols (ICNP)*, 2006.

[30] Kastenholz F, ISLAY:A new routing and addressing architecture, *IRTF, Internet Draft*, 2002.

[31] Farinacci D, Fuller V, Oran D, and Meyer D, Locator/ID Separation Protocol(LISP), *Internet Draft*, *http://www.ietf.org/internet-drafts/draft-farinacci-lisp-05.txt*, 2007.

[32] Whittle, R, Ivip (Internet Vastly Improved Plumbing) Architecture, *Internet Draft, Work in Progress, draft-whittle-ivip-arch-01.txt.*

[33] Vogt C, Six/One Router: A scalable and backwards-compatible solution for provider-independent addressing, In *ACM SIGCOMM MobiArch Workshop*, 2008.

[34] Yang X, Clark D, Berger AW, NIRA: A new inter-domain routing architecture, *IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 15, NO. 4*, AUGUST 2007.

[35] Godfrey PB, Ganichev I, Shenker S, Stoica I, Pathlet routing, In *Proceedings of SIGCOMM*, Aug. 2009.

作者简介：

**唐明董：** 中国科学院计算技术研究所博士生，湖南科技大学计算机学院讲师，Email: tangmingdong@ict.ac.cn

**张国清：** 博士，中国科学院计算技术研究所副研究员，硕士生导师

**杨　景：** 博士，中国移动研究院首席科学家，中国科学院计算技术研究所客座研究员，博士生导师

**张国强：** 博士，中国科学院计算技术研究所助理研究员